

Code Generation: Results Achieved with Zero-shot Prompting When Asking for a 5-level Judgment

LLM	Boolean judgment		5-level judgment	
	#	%	#	%
DeepSeek Coder 1.3B	687	42.54%	39	2.41%
DeepSeek Coder 6.7B	22	1.36%	2	0.12%
DeepSeek Coder 33B	4	0.25%	2	0.12%
CodeLlama 7B	1	0.06%	1	0.06%
CodeLlama 13B	3	0.19%	0	0.00%
CodeLlama 34B	2	0.12%	2	0.12%
GPT-3.5-turbo	0	0.00%	0	0.00%
GPT-4-turbo	27	1.67%	0	0.00%

Table 1: Number (#) and percentage (%) of instances for which the LLMs did not manage to output a valid judgment. Comparison between the boolean and the 5-level scenarios.

DSC 1.3B	DSC 6.7B	DSC 33B	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4
0.02	-0.02	0.08	-0.13	0.05	0.24	0.15	0.29

Table 2: Code Generation: Spearman correlation coefficient between the 5-level judgments of the LLMs and the pass/fail (0 or 1) ground truth. The Kappa agreement is not shown because the judgments of the LLMs range from 1 to 5, whereas the ground truth is binary.

	DSC 1.3B	DSC 6.7B	DSC 33B	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4	Human Written	Own vs LLMs	Own vs LLMs \ F	Own vs Human
DSC 1.3B	3.95	4.00	4.00	4.00	3.97	4.00	3.98	4.00	3.91	(N)	(N)	(N)
DSC 6.7B	5.00	4.98	4.91	4.98	4.97	4.97	5.00	5.00	4.68	(N)	(N)	*** (S)
DSC 33B	4.82	4.87	4.85	4.79	4.89	4.90	4.89	4.93	4.77	(N)	(N)	(N)
CL 7B	4.58	4.49	4.07	4.49	4.37	4.52	4.35	4.33	4.06	(N)	* (S)	*** (M)
CL 13B	3.98	3.98	3.97	3.98	3.95	3.95	4.00	3.99	4.01	(N)	(N)	(N)
CL 34B	3.76	4.17	3.76	4.07	3.95	3.61	4.18	4.16	3.34	* (S)	* (S)	(N)
GPT-3.5	4.78	4.87	4.70	4.72	4.77	4.75	4.89	4.87	4.04	(N)	(N)	*** (M)
GPT-4	4.58	4.81	4.61	4.47	4.71	4.46	4.93	4.97	3.57	** (S)	** (S)	*** (L)
Average (all)	4.43	4.52	4.36	4.44	4.45	4.40	4.53	4.53	4.05	-	-	-
Average (large)	4.38	4.54	4.38	4.41	4.46	4.33	4.58	4.58	3.95	-	-	-

Adjusted p -values: * <0.05, ** <0.01, *** <0.001. Cliff delta: N=Negligible, S=Small, M=Medium, L=Large

Table 3: 5-level scenario: average rating given by the judge LLM (rows) to the functions generated by the generator LLM or manually written by humans (columns). Note that here only functions **passing** the tests are considered. Last three columns report adj. p -value and effect size when comparing the judgements each LLM gave to functions it generated against those it gave when judging functions (i) generated by all other LLMs, (ii) generated by all other LLMs but those belonging to the same family, and (iii) written by humans.

	DSC 1.3B	DSC 6.7B	DSC 33B	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4	Own vs LLMs	Own vs LLMs \ F
DSC 1.3B	3.93	3.97	3.96	3.95	3.94	3.92	4.06	3.92	(N)	(N)
DSC 6.7B	4.91	4.97	4.73	4.96	4.95	4.93	4.81	4.95	(N)	(N)
DSC 33B	4.64	4.69	4.77	4.71	4.78	4.78	4.82	4.83	(N)	(N)
CL 7B	4.48	4.51	4.10	4.43	4.37	4.50	4.47	4.36	(N)	(N)
CL 13B	3.93	3.96	3.91	3.96	3.96	3.91	4.00	3.97	(N)	(N)
CL 34B	3.05	3.63	3.07	3.60	3.12	3.16	3.28	3.25	(N)	(N)
GPT-3.5	3.88	4.28	4.06	4.19	4.32	4.22	4.29	4.56	(N)	(N)
GPT-4	3.17	3.67	3.79	3.52	3.61	3.36	3.50	4.39	*** (S)	*** (M)
Average (all)	4.00	4.21	4.05	4.17	4.13	4.10	4.15	4.28	-	-
Average (large)	3.73	4.04	3.92	4.00	3.96	3.89	3.98	4.20	-	-

Adjusted p -values: * <0.05, ** <0.01, *** <0.001. Cliff delta: N=Negligible, S=Small, M=Medium, L=Large

Table 4: 5-level scenario: average rating given by the judge LLM (rows) to the functions generated by the generator LLM or manually written by humans (columns). Note that here only functions **failing** the tests are considered. The *Human Written* column is not present because there are no human written function which fail the tests. Last three columns report adj. p -value and effect size when comparing the judgements each LLM gave to functions it generated against those it gave when judging functions (i) generated by all other LLMs, (ii) generated by all other LLMs but those belonging to the same family, and (iii) written by humans.

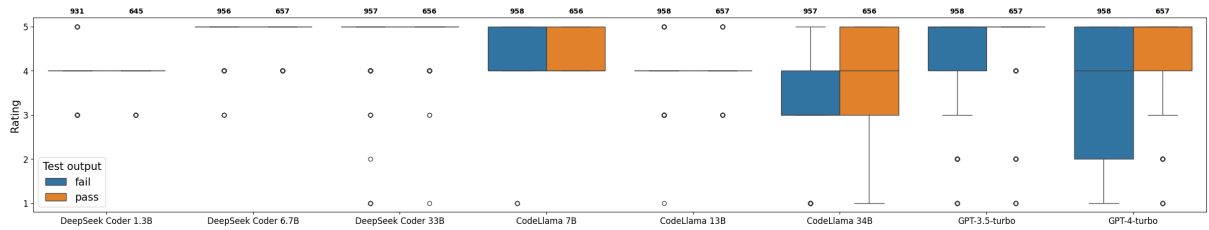


Figure 1: Code Generation: boxplots of judgments provided by the 8 LLMs in the 5-level scenario.